SPSS

Statistics for Data Analysis

WHO WE ARE

SPS is an Italian center of statistical data analysis with more than 20 years of experience.

SPS was born in 1994 as SPSS Italia and it was the only reseller in Italy for SPSS software suite, authorised by SPSS inc.

Today SPS is an IBM Gold Business Partner, Software Support Provider and Expert Level in Data Science & Business Analytics.

CONTACTS

Registered office:
Via Antonio Zanolini, 36 A/B
40126 Bologna (BO

Operational headquarters:
Via Isonzo, 55/2
40033 Casalecchio di Reno (BO)
P.I. 04222630370
Tel. 051-252573
www.spss.it

DATASHEET

Data Preparation

Statistics for Data Analysis

Highlights Statistics for Data Analysis

- Get support through every step of the analytical process.
- Carry out essential analyses from an intuitive graphical interface.
- Select from more than a dozen integrated products to make specialized analyses faster and easier.

Organizations can solve a wide array of business and research problems with the solution Statistics for Data Analysis.

Compared to other statistical software, the solution is easier to use, has a lower total cost of ownership and more comprehensively addresses the entire analytical process, from planning to data collection to analysis, reporting and deployment.

Organizations of all types rely on Statistics for Data Analysis to help increase revenue, outmaneuver competitors, conduct research and make better decisions. With decades of built-in expertise and innovation, it's a leading choice for reliable statistical analysis.

Statistics Base is part of the solution Statistics for Data Analysis, which consists of:

- Software license
- Add-On
- SPS Service Program

This comprehensive, easy-to-use solution includes many different procedures and tests to help users solve complex business and research challenges.

Statistics for Data Analysis

The solution analytical capabilities to meet the analysis requirements of any type of organization, from basic tools for solving common problems to advanced analytical techniques that enable all type of organization to address complex challenges.

Statistics for Data Analysis can help you:

- Analyze your data with new and advanced statistics, including a variety of new features within UNIANOVA methods
- Integrate better with third-party applications, including stronger integration with Microsoft Office
- Save time and effort with productivity enhancements:
  - More attractive and modern-looking charts in Chartbuilder
  - New groundbreaking features in Statistics Amos 25
  - Data and syntax editor enhancements
  - Accessibility improvements for the visually impaired
  - Updated merge user interface
  - Simplified toolbars

Statistics for Data Analysis can access quickly, manage and analyze any kind of dataset, including survey data, corporate databases or data downloaded from the web.

In addition, the software can process Unicode data. This eliminates variability in data due to language-specific encoding and enables your organization to view, analyze and share data written in multiple languages.

Business Benefit Statistics for Data Analysis

- Support business decisions with data-based analytics for improved outcomes.
- Be more confident in your results by incorporating data from many different sources, including geospatial information, in your analysis and using proven, tested techniques to perform your analysis.
- Save time and effort with capabilities that enable experienced analysts to develop procedures or dialogs that others can use to speed through repetitive tasks.
- Give results greater impact by using visualization capabilities that clearly show others the significance of your findings.

Statistics Data Preparation

Datasheet

Improve data preparation for more accurate results

Highlights:

- Identify suspicious or invalid cases, variables, and data values.
- View patterns of missing data.
- Summarize variable distributions.
- Prepare data for analysis more accurately, quickly.

All researchers have to prepare their data prior to analysis. While Statistics includes tools for data preparation, sometimes you need more specialized techniques to get your data ready. With Statistics Data Preparation, you can easily identify suspicious or invalid cases, variables and data values; view patterns of missing data; summarize variable distributions; and more accurately work with algorithms designed for nominal attributes. This streamlines the data preparation process—so that you can get ready for analysis faster and reach more accurate conclusions. Choose from a completely automated data preparation procedure for the fastest results, or select from several other methods to help you handle more challenging datasets.

Choose from several options for data preparation

The Validate Data procedure

Data validation has typically been a manual process. You might run a frequency on your data, print the frequencies, circle what needs to be fixed and check for case IDs. This is time consuming and, since every analyst in your organization could use a slightly different method, maintaining consistency from project to project may be a challenge.

Statistics Data Preparation

Datasheet

To eliminate manual checks, use the Validate Data procedure. This procedure enables you to apply rules to perform data checks based on each variable's measure level (whether categorical or continuous). For example, if you're analyzing survey data that has variables on a five-point Likert scale, use the Validate Data procedure to apply a rule for five-point scales and flag all cases that have values outside of the 1-5 range. You can receive reports of invalid cases as well as summaries of rule violations and the number of cases affected, as well as specify validation rules for individual variables (such as range checks) and cross-variable checks (for example, "pregnant males").
This knowledge can help you determine data validity and remove or correct suspicious cases at your discretion prior to analysis.

Prepare data in a single step, automatically

Manual data preparation is a complex process that can account for as much as 40 to 90 percent of an analyst's time on a given project. When you need results quickly, the Automated Data Preparation (ADP) procedure helps you detect and correct quality errors and impute missing values in one efficient step. The ADP feature provides an easy-to-understand report with complete recommendations and visualizations to help you determine which data to use in your analysis.

Statistics Data Preparation

Datasheet

Optimal Binning

In order to use algorithms that are designed for nominal attributes (such as Naïve Bayes and logit models), you must bin your scale variables before model building. If scale variables aren't binned, algorithms such as multinominal logistic regression will take an extremely long time to process, or they might not converge, especially if you have a large dataset. In addition, the results you receive may be difficult to read or interpret.

Optimal Binning, however, enables you to determine cutpoints to help you reach the best possible outcome for algorithms designed for nominal attributes.

With this procedure, you can select from three types of binning for preprocessing data prior to model building:

- Unsupervised – Create bins with equal counts
- Supervised – Take the target variable into account to determine cutpoints. This method is more accurate than unsupervised; however, it is also more computationally intensive.
- Hybridapproach – Combines the unsupervised and supervised approaches. This method is particularly useful if you have a large amount of distinct values.

Statistics Data Preparation Features

Automated Data Preparation
Recommend steps to speed up model building and improve predictive power:

- Determine Objective: Balance speed and accuracy, Optimize for speed, Optimize for accuracy, or Customize analysis
- Prepare dates and times for modeling:
  - Compute elapsed time until a reference date
  - Compute elapsed time until a reference time
  - Extract cyclical time elements
- Exclude low-quality input fields:
  - Exclude fields with too many missing values
  - Exclude nominal fields with too many unique categories
  - Exclude categorical fields with too many values in a single category
- Adjust measurement levels:
  - Adjust measurement levels of numeric fields
- Prepare fields to improve data quality:
  - Outlier handling
  - Replace missing values
  - Reorder nominal fields
- Rescale Fields:
  - Analysis weight
- Continuous input fields
- Continuous target fields
- Transform Fields:
  - Using both categorical and/or continuous input fields
- Perform feature selection and construction
- Name fields:
  - Transformed and constructed fields
  - Computed durations
  - Extracted cyclical time elements
- Apply transformations to data

Validate data
Use the Validate Data procedure to validate data in the working data file: Basic checks: Specify basic checks to apply to variables and cases in your file.

- For example, obtain reports that identify variables with a high percentage of missing values or empty cases:
  - Maximum percentage of missing values
  - Maximum percentage of cases in a single category
  - Maximum percentage of cases with a count of 1
  - Minimum coefficient of variation
  - Minimum standard deviation
  - Flag incomplete IDs
  - Flag duplicate IDs
  - Flag empty cases
- Standard rules: Describe the data, view single variable rules and apply them to analysis variables:
  - Description of data:
    - Distribution: Shows a thumbnail-size bar chart for categorical variables or histogram for scale variables
    - Minimum and maximum data values are shown
  - Single-variable rules:
    - Apply rules to individual variables to identify missing or invalid values, such as values outside a valid range
    - User-defined single-variable rules are also possible

- Custom rules: Define cross-variable rule expressions in which respondents' answers violate logic ("pregnant males", for example)
- Output: Reports describing invalid data:
  - Casewise report, which lists the validation rule violations by case:
    - Specify the minimum number of violations needed for a case to be included in the report
    - Specify the maximum number of cases in the report
  - Standard validation rules reports:
    - Summarize violations by analysis variable
    - Summarize violations by rule
    - Display descriptive statistics
- Save: Enables you to save variables that record rule violations and use them to help clean data and filter out bad cases:
  - Summary variables:
    - Empty case indicator
    - Duplicate ID indicator
    - Incomplete ID indicator
    - Validation rule violation (total count)
  - Indicator variables that record all validation rule violations

Identify unusual cases

The Anomaly Detection procedure searches for unusual cases, based upon deviations from their peer group, and gives reasons for such deviations:

- Specify variables to be used by the procedure with the VARIABLES subcommand. Specify categorical, continuous, and ID variables (to identify cases), and list variables that are excluded from the analysis.
- The HANDLEMISSING subcommand specifies the methods of handling missing values in this procedure:
  - Apply missing value handling. If this option is selected, grand means are substituted for missing values of continuous variables, and missing categories of categorical variables are combined and treated as a valid category. The processed variables are then used in the analysis. If this option is not selected, cases with missing values are excluded from the analysis.
  - Create an additional Missing Proportion Variable and use it in the analysis. If chosen, an additional variable called the Missing Proportion Variable that represents the proportion of missing variables in each record is created, and this variable is used in the analysis. If it is not chosen, the Missing Proportion Variable is not created.
- The CRITERIA subcommand specifies the following settings:
  - Minimum and maximum number of peer groups
  - Adjustment weight on the measurement level
  - Number of reasons in the anomaly list
  - Percentage of cases considered as anomalies and included in the anomaly list
  - Number of cases considered as anomalies and included in the anomaly list
  - Cutpoint of the anomaly index to determine whether a case is considered as an anomaly

- Save additional variables to the working data file with the SAVE subcommand:
  - Anomaly index
  - Peer group ID
  - Peer group size
  - Peer group size in percentage
  - The variable, associated with a reason
  - The variable impact measure, associated with a reason
  - The variable value, associated with a reason
  - The norm value, associated with a reason
- Write the model to a specified filename as XML with the OUTFILE subcommand
- Control the display of the output results with the PRINT subcommand
- You can print:
  - Case-processing summary
  - The anomaly index list, the anomaly peer ID list and the anomaly reason list
  - The Continuous Variable Norms table, if any continuous variable is used in the analysis, and the Categorical
  - Variable Norms, if any categorical vari able is used in the analysis
  - Anomaly Index Summary
  - Reason Summary Table for each reason:
    - Suppress all displayed output except the notes table and any warnings

Optimal Binning

Preprocess data using Optimal Binning, which categorizes one or more continuous variables by distributing the values of each variable into bins. This procedure is useful for reducing the number of values in the given binning input variables, which can greatly improve the performance of algorithms. When using certain Optimal Binning methods, a guide variable helps you determine the cutpoints, thereby maximizing the relationship between the guide variable and the binned variable.

- Select from the following methods:
  - Unsupervised binning via the equal frequency algorithm. This method uses the equal frequency algorithm to discretize the binning input variables. A guide variable is not required.
  - Supervised binning via the MDLP (Minimal Description Length Principle) algorithm. This method discretizes the binning input variables using the MDLP algorithm without any preprocessing. It is suitable for datasets with a small number of cases. A guide variable is required.
  - Hybrid MDLP binning. This involves preprocessing via the equal frequency algorithm, followed by the MDLP algorithm. This method is suitable for datasets with a large number of cases. A guide variable is required.

- Specify the following criteria:
  - How to define the minimum cutpoint for each binning input variable
  - How to define the maximum cutpoint for each binning input variable
  - How to define the lower limit of an interval
  - Whether to force merging of sparsely populated bins
  - Whether missing values are handled using listwise or pairwise deletion
- Save the following:
  - New variables containing binned values
  - Syntax to an Statistics Base syntax file
- Control output results display with the PRINT subcommand. You can print:
  - --The binning input variables' cutpoint sets
  - --Descriptive information for all binning input variables
  - --Model entropy for binned variables

Statistics for Data Analysis solution

Add more analytical power, as you need it, with optional modules and stand-alone software from the Statistics for Data Analysis family.

Statistics Base

Statistics Base includes the core capabilities to take the analytical process from start to finish. It is easy to use and includes a broad range of procedures and techniques to increase revenue, outperform competitors, conduct research and make better decisions.

Statistics Advanced

Statistics Advanced includes these powerful multivariate techniques: generalized linear models (GENLIN), generalized estimating equations (GEE), mixed level models, general linear mixed models (GLMM), variance component estimation, MANOVA, Kaplan-Meier estimation, Cox regression, hiloglinear, loglinear and survival analysis.

Statistics Bootstrapping

Statistics Bootstrapping enables researchers and analysts to use bootstrapping techniques on a number of tests contained in Statistics for Data Analysis modules. This provides an efficient way to ensure that your models are stable and reliable. With Statistics Bootstrapping, you can reliably estimate the standard errors and confidence intervals of a population parameter like a mean, median, proportion, odds ratio, correlation coefficient, regression coefficient and numerous.

Statistics Categories

Unleash the full potential of your categorical data through perceptual maps with optimal scaling and dimension reduction techniques. This add-on module provides you with everything you need to analyze and interpret multivariate data and their relationships more completely.

Statistics Complex Samples

Incorporate complex sample designs into data analysis for more accurate analysis of complex sample data. Statistics Complex Samples, with specialized planning tools and statistics, reduces the risk of reaching incorrect or misleading inferences for stratified, clustered or multistage sampling.

Statistics Conjoint

Statistics Conjoint helps market researchers develop successful products. By performing conjoint analysis, you learn what product attributes are important in the consumer's mind and what the most preferred attribute levels are, and can perform pricing studies and brand equity studies.

Statistics Tables

Use Statistics Tables to present survey, customer satisfaction, polling and compliance reporting results. Features such as a table builder preview, included inferential statistics and data management capabilities make it easy to clearly communicate your results.

Statistics Preparation
With Statistics Preparation, you gain several procedures that facilitate the data preparation process. This add-on module enables you to easily identify suspicious and invalid cases, variables and data values; view patterns of missing data; summarize variable distributions to get your data ready for analysis; and more accurately work with algorithms designed for nominal attributes.

Statistics Decision Trees
Create highly visual classification and decision trees directly within Statistics for Data Analysis for segmentation, stratification, prediction, data reduction and variable screening, interaction identification, category merging and discretizing continuous variables. Highly visual trees enable you to present results in an intuitive manner.

Statistics Direct Marketing
Statistics Direct Marketing helps marketers perform various kinds of analyses easily and confidently, without requiring a detailed understanding of statistics. They can conduct recency, frequency and monetary value (RFM) analysis, cluster analysis, and prospect profiling. They can also improve marketing campaigns through postal code analysis, propensity scoring, and control package testing. And they can easily score new customer data and access pre-built models.

Statistics Exact Tests
Statistics Exact Tests always provides you with correct p values, regardless of your data structure, even if you have a small number of cases, have subset your data into fine breakdowns or have variables where 80 percent or more of the responses are in one category.

Statistics Forecasting
Improve forecasting with complete time-series analyses, including multiple curve-fitting, smoothing models, methods for estimating autoregressive functions and temporal causal modeling. Use the Expert Modeler to automatically determine

which ARIMA (autoregressive integrated moving average) process or exponential smoothing model best fits your time-series and independent variables, eliminating selection through trial and error.

Statistics Missing Values
If values are missing from your data, this module may find some relationships between the missing values and other variables. In addition, the missing values module can estimate what the value would be if data weren't missing.

Statistics Neural Networks
Use the Statistics Neural Networks module to model complex relationships between inputs and outputs or to discover patterns in your data. Choose from algorithms that can be used for classification (categorical outcomes) and prediction (numerical outcomes). The two available algorithms are Multilayer Perceptron and Radial Basis Function.

Statistics Regression
Predict behavior or events when your data go beyond the assumptions of linear regression techniques. Perform multinomial or binary logistic regression and nonlinear regression, weighted least squares, two-stage least squares and probit analysis.

Complementary product
Use these products with Statistics for Data Analysis to enhance your analytical results.

Statistics Amos
Support your research and theories by extending standard multivariate analysis methods when using this stand-alone software package for structural equation modeling (SEM). Build attitudinal and behavioral models that more realistically reflect complex relationships, because any numeric variable, whether observed or latent, can be used to predict any other numeric variable. The latest release includes a new nongraphical method of model specification that improves accessibility for users who need scripting capabilities and enables large, complicated models to be run more quickly.